Shapley explainability on the data manifold

Year: 2020 | Citations: 113 | Authors: Christopher Frye, Damien de Mijolla, T. Begley, Laurence Cowton, Megan Stanley

Abstract

Explainability in AI is crucial for model development, compliance with regulation, and providing operational nuance to predictions. The Shapley framework for explainability attributes a model's predictions to its input features in a mathematically principled and model-agnostic way. However, general implementations of Shapley explainability make an untenable assumption: that the model's features are uncorrelated. In this work, we demonstrate unambiguous drawbacks of this assumption and develop two solutions to Shapley explainability that respect the data manifold. One solution, based on generative modelling, provides flexible access to data imputations; the other directly learns the Shapley value-function, providing performance and stability at the cost of flexibility. While "off-manifold" Shapley values can (i) give rise to incorrect explanations, (ii) hide implicit model dependence on sensitive attributes, and (iii) lead to unintelligible explanations in higher-dimensional data, on-manifold explainability overcomes these problems.